



# Some Basic Concepts of Survey Sampling

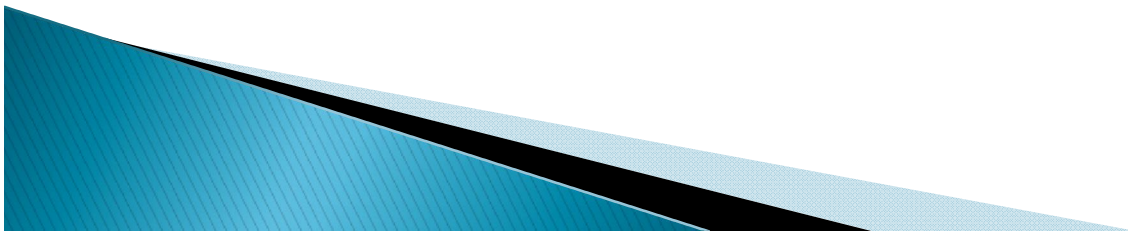
By:

**Prof. Sheela Misra**  
**Former Head, Dept of Statistics**  
**University of Lucknow, Lucknow**

[Email:profsheelamisra@gmail.com](mailto:profsheelamisra@gmail.com)

# Some quotes

- ▶ To do successful research you don't need to know everything; you just need to know one thing that isn't known
- ▶ Great design is great complexity presented via simplicity
- ▶ It's called a research. It's how people install a new software in their brains

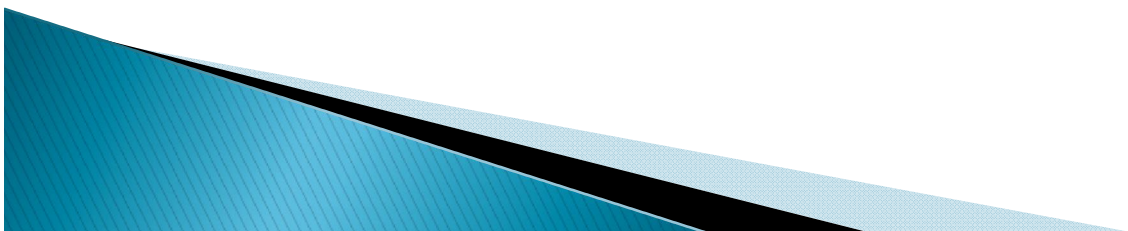


Sequence of activities in any research endeavour in the correct order consist of:

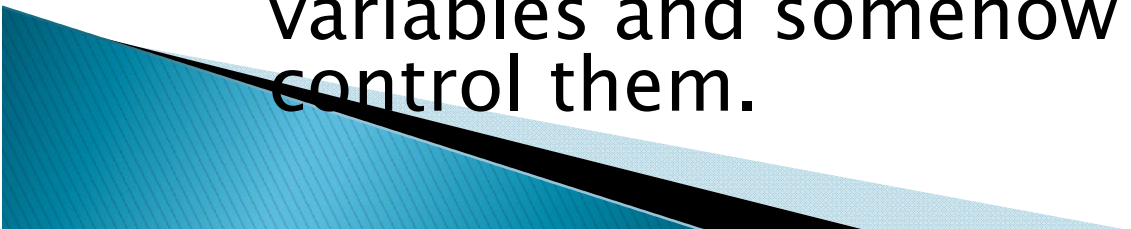
- Selecting
- Observing
- Recording
- Comparing
- Analyzing
- Classifying

The Phenomenon  
of interest

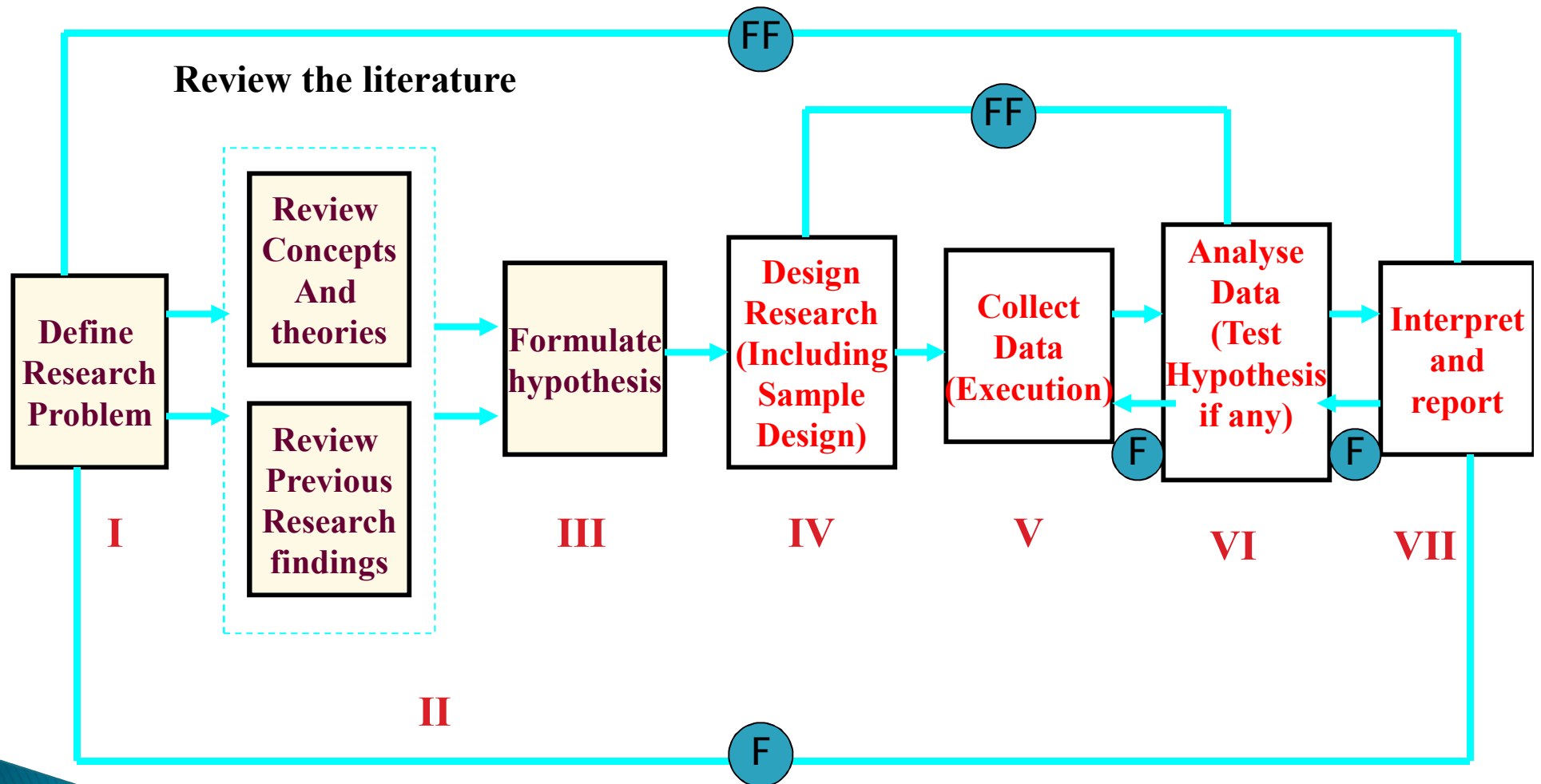
- ▶ Deriving appropriate inference(s) & attempting integration &
- ▶ Publication



# Variables

- ▶ Researchers measure variables.
  - ▶ Qualitative and Quantitative –temperature
  - ▶ For quantitative research we distinguish
  - ▶ The independent variable (the cause), while the dependent variable is the effect (or assumed effect), –important in experimental research.
  - ▶ Confounding variables are variables with a significant effect on the dependent variable that the researcher failed to control or eliminate – sometimes the researcher is not aware of the effect of the confounding variable.
  - ▶ The key is to identify possible confounding variables and somehow try to eliminate or control them.
- 

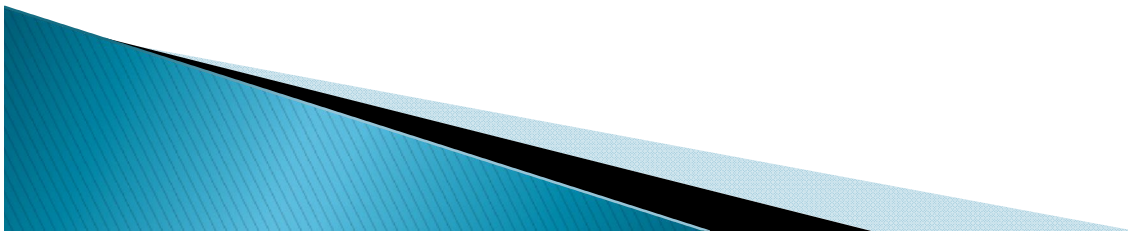
# RESEARCH PROCESS



**F** — Feed Back  
**FF** — Feed Forward

# Quantitative methods


- ▶ You can approach statistical work with any or all of the following steps:
- ▶ Identifying population and sample
- ▶ ■ Framing your Research Methodology
- ▶ ■ Study design
- ▶ ■ Sample size calculation and justification
- ▶ ■ Development of questionnaire
- ▶ ■ Statistical techniques
- ▶ Drawing conclusions



# Populations and Samples

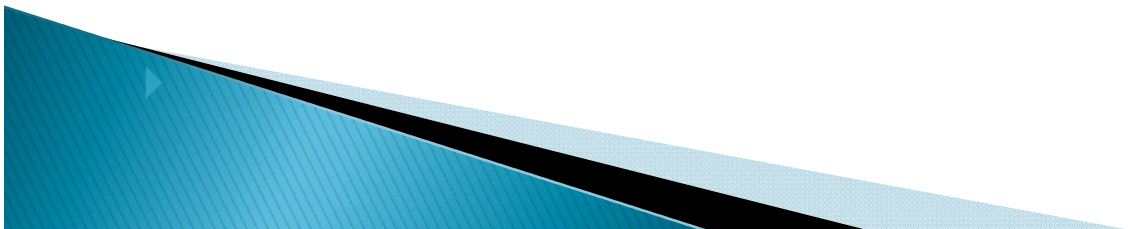
- ▶ A **population** –any set of persons/subjects having a common observable characteristic; large/ small/ finite/ infinite; all pregnant women make up a population. The characteristics of a population–parameter
- ▶ A **sample** – a subset of the population. The characteristics of a sample – a statistic.
- ▶ Why Sample? four major reasons to sample.
- ▶ First, usually too costly to test the entire population.; Census every ten years.
- ▶ The second reason –it may be impossible to test the entire population.
- ▶ The third reason to sample is that testing the entire population often produces error. Thus, sampling may be more accurate.
- ▶ Rating of patients in case of depression– consistency from one patient to the next; many raters introduce a source of error;. different criteria .
- ▶ The final reason to sample is that testing may be destructive.
- ▶ It makes no sense to lesion the lateral hypothalamus of all rats to determine if it has an effect on food intake.
- ▶ We can get that information from operating on a small sample of rats. Also, you probably would not want to buy a car that had the door slammed five hundred thousand time or had been crash tested.


# Sampling Procedures

- ▶ A sample consists of a subset of the population. Any member of the defined population can be included in a sample.
  - ▶ A theoretical list (an actual list may not exist) of individuals or elements who make up a population is called a sampling frame; five major sampling procedures.
  - ▶ The first sampling procedure is convenience but not reliable– Volunteers, members of a class, individuals in the hospital with the specific diagnosis biased.
  - ▶ We consider only random sampling methods.
  - ▶ Simplest is the simple random sampling – all subjects or elements have an equal chance of being selected.
- 



- ▶ A systematic sample is conducted by randomly selecting a first case from list of the population and then proceeding every kth, say 10<sup>th</sup> case until your sample is selected. Long list
- ▶ For example, if your list was the phone book, it would be easiest to start at perhaps the 17<sup>th</sup> person, and then select every 50<sup>th</sup> person from that point on.  $K = N/n$
- ▶ Stratified sampling.  
In a stratified sample, we sample either proportionately or equally to represent various strata or subpopulations.
- ▶ For example if our strata were states we would make sure all the fifty states represented.  
 $N = \text{strata}$
- ▶ Strata religious affiliation, rural/urban
- ▶  $n_h \text{ prop } N_h \text{ or } n_h \text{ prop } N_h S_h$



- ▶ Cluster sampling
  - ▶ In cluster sampling we take a random sample of clusters and then survey every member of the group. For example, if our cluster were children in schools in the City Montessori Public Schools, we would randomly select perhaps 5 schools and then test all of the students within those schools.
  
  - ▶ Sampling Problems
  - ▶ There are several potential sampling problems. When designing a study, a sampling procedure is also developed including the potential sampling frame. Several problems may exist within the sampling frame.
  - ▶ First, there may be missing elements—unlisted no.
  - ▶ Foreign elements— Elements which should not be included in my population and sample appear on my sampling list.
  - ▶ Duplicates – These are elements who appear more than once on the sampling frame.
- 

# 30 cluster sampling

- ▶ This method is the most frequently used in the field in the large population based study –two stages.
- ▶
- ▶ In first stage the entire population is divided into small distinct geographic areas, such as villages, camps, etc. Find size
- ▶ In second stage, the random selection of households within clusters are chosen randomly within each cluster using simple or systematic random sampling.
- ▶ The sampling interval is calculated as the total population of all the geographic units divided by the number of clusters needed.
- ▶ Population=4200, clusters=30, sampling interval= $4200/30=140$
- ▶ Select one no. 1–140, say 100, falls in a village–the first cluster then  $100+140=240^{\text{th}}$  village (cluster).....all 30 clusters

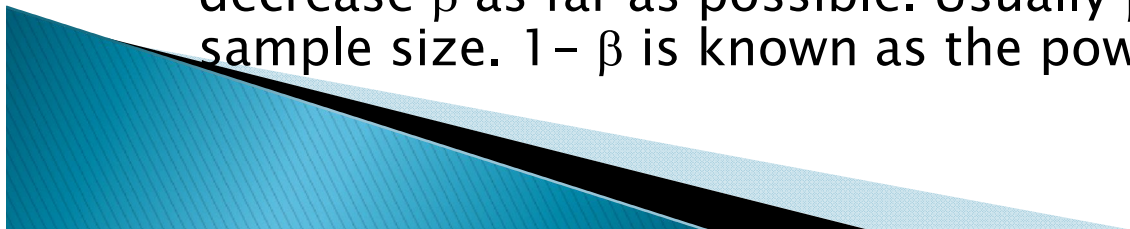
▶ Type-I and Type-II Errors in Decision Making

▶ Decision by the investigator Existing Reality


Decision	Drug B=Drug A Type-I Error ( $\alpha$ )	Drug B#Drug A Correct (power; $1 - \beta$ )
Drug B= Drug A	Correct Decision (Level of confidence $1 - \alpha$ )	Type-II Error ( $\beta$ )

In any study, no investigator can be 100% sure that the decision taken is correct.

Attempt is to take  $\alpha$  as minimum at a fixed level(0.05) and then decrease  $\beta$  as far as possible. Usually  $\beta=0.10$  taken for estimating sample size.  $1 - \beta$  is known as the power of the test.

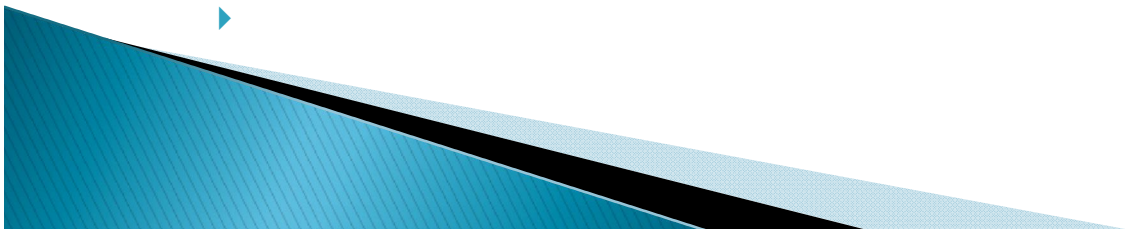


# Sample size

- ▶ Too small a sample is one of the most common problems seen in clinical trials, testing significance of difference
  - ▶ Sample size goes hand-in-hand with power: if no difference between treatments is seen, it could be because there was not enough power to detect a true difference.
  - ▶ This will decrease the standard error and hence, increase our ability to detect true treatment difference.
  - ▶ In absence of minimum required sample size, the experiment has a poor chance of detecting important treatment differences and would simply be a waste of time and money.
- 

# Sample size

- ▶ If data is not readily available for the process, what should be the sample size so the population is properly represented?
- ▶ If data has been collected, how do you determine if you have enough data?
- ▶ Determining sample size is a very important issue because too large samples may waste time, resources and money, while samples too small may lead to inaccurate results
- ▶ We would like to start an Internet Service Provider (ISP) and need to estimate the average Internet usage of households in one week for our business plan and model.
- ▶ How many households must we randomly select to be 95% sure that the sample mean is within 1 minute of the population mean? .
- ▶ Assume that a previous survey of household usage has shown standard deviation  $\sigma = 6.95$  minutes.
- ▶



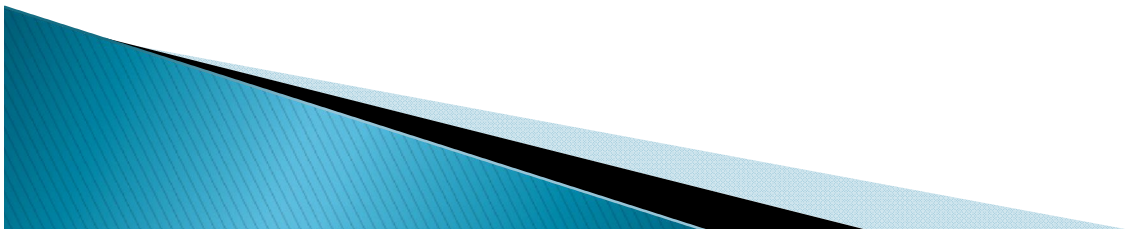
- ▶ A 95% degree confidence corresponds to level of significance = 0.05. Each of the shaded tails in the normal curve has an area of = 0.025. The region to the left of and to the right of = 0 is 0.5 - 0.025, or 0.475 on each side.
- ▶ In the Table of the Standard Normal Distribution, an area of 0.95 corresponds to a z value of 1.96. The critical value is therefore = 1.96.
- ▶ E=margin of error–difference between sample mean and expected population mean
- ▶  $n = \frac{(1.96 \text{ s.d.})^2}{E^2}$
- ▶  $n = \frac{4 \text{ s.d.}^2}{E^2}$  or  $4PQ / E^2$  (large pop.)

- ▶ Let us calculate the sample size .
- ▶
- ▶ The margin of error  $E = 1$  and the standard deviation  $\sigma = 6.95$ . Using the formula for sample size, we can calculate :



$$n = \left[ \frac{z_{\alpha/2} \cdot \sigma}{E} \right]^2 = \left[ \frac{1.96 \cdot 6.95}{1} \right]^2 = [13.62]^2 = 185.55 = 186$$

- ▶ So we randomly selected 186 households. With this sample we will be 95% confident that the sample mean will be within 1 minute of the true population of Internet usage.






- ▶ Higher confidence level requires a larger sample size.
- ▶ What is the population size? If you don't know, use 20000
- ▶ The sample size doesn't change much for populations larger than 20,000.
- ▶ What is the response distribution? For each question, what do you expect the results will be?
- ▶ If the sample is skewed highly one way or the other, the population probably is, too.
- ▶ If you don't know, use 50%, which gives the largest sample size.  $n_0 = 385 \quad (1.96 \text{ s.d.})^2 / E^2$
- ▶ Your recommended sample size is 377– the minimum recommended size of your survey.
- ▶ If you create a sample of this many people and get responses from everyone, you're more likely to get a correct answer than you would from a large sample where only a small percentage of the sample responds to your survey.



# “ What affects size of sample?”

- ▶ Influenced by a number of factors, including the purpose of the study, population size, the risk of selecting a "bad" sample, and the allowable sampling error.
  - ▶ In addition to the purpose of the study and population size, three criteria usually will need to be specified to determine the appropriate sample size: *the level of precision, the level of confidence or risk, and the degree of variability*
  - ▶ The Level of Precision  
The *level of precision*, sometimes called *sampling error*, is the range in which the true value of the population is estimated to be (e.g.,  $\pm 5$  percent) in the same way that results for political campaign polls are reported by the media.
  - ▶ Thus, if a researcher finds that 60% of farmers in the sample have adopted a recommended practice with a precision rate of  $\pm 5\%$ , then he or she can conclude that between 55% and 65% of farmers in the population have adopted the practice.
- 

Thank You

